VU Research Portal

A tool for gene expression based PubMed search through combining data sources.

Korotkiy, M.; Middelburg, R.A.; Dekker, H.; van Harmelen, F.A.H.; Lankelma, J.

published in
Bioinformatics
2004

DOI (link to publisher)
10.1093/bioinformatics/bth183

document version
Publisher's PDF, also known as Version of record

BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 12 2004, pages 1980–1982
doi:10.1093/bioinformatics/bth183

A tool for gene expression based PubMed search through combining data sources

Maksym Korotkiy[1], Rutger Middelburg[2], Henk Dekker[2], Frank van Harmelen[1] and Jan Lankelma[2,3,*]

[1]Division of Computer Science, [2]Department of Medical Oncology, VU University Medical Center and [3]Tumor Cell Biology Group, Faculty of Earth and Life Sciences, Vrije Universiteit, De Boelelaan 1087, 1081 HV Amsterdam, The Netherlands

Received on August 28, 2003; revised on March 4, 2004; accepted on March 10, 2004
Advance Access publication March 25, 2004

ABSTRACT

Summary: We present a new tool for the semi-automated querying of PubMed using a batch of tens to thousands of GenBank accession numbers or UniGene cluster ids. By combining information from UniGene and SWISS-PROT, microGENIE obtains information on the biological relevance of expressed genes, as identified by micro-array experiments, with minimal user intervention and time investment.

Availability: microGENIE is freely available from http://www.cs.vu.nl/microgenie

Contact: jan.lankelma@falw.vu.nl

Supplementary information: The web site above supplies examples of input and output files.

INTRODUCTION

In recent years, the use of micro-arrays has expanded dramatically. Especially, the use of mRNA expression-arrays is becoming a standard technique. Major difficulties in the experimental procedures have been resolved and the data extraction has also received the appropriate attention.

However, the analysis of the obtained data still poses a problem. This data consists of the relative levels of mRNA expression of thousands of genes compared between two or more different samples. The unveiling of the biological relevance of these expression profiles still offers a challenge that is generally met by the manual screening of the literature for relevant information.

Usually only a subset of the micro-array data is considered to be interesting for further analysis. Thus, the amount of data to be analyzed is greatly reduced but there will often be several hundreds of accession numbers left for a literature search. Unfortunately, accession numbers[1] are not used in publications. Therefore, a researcher will have to determine which gene is represented by a given spot on the micro-array, before it is possible to search for relevant literature.

PubMed (http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html) is a publicly available electronic library of publications in the field of biomedical research. The National Center for Biotechnology Information (NCBI) maintains the library and provides search capabilities that allow querying of the library.

After each experiment the following steps have to be performed for every spot of interest:

(1) find out what gene the spot is associated with;
(2) craft a PubMed query for publications about this gene; and
(3) analyze these publications.

Steps 1 and 2 are semi-automated in our tool requiring only minor intervention from the user. Assisting the user with the further analysis of the obtained publications is not yet addressed in our current work.

A few analytical tools are available that can be employed to automate the process of extracting biologically relevant information from the micro-array data (Tanabe *et al*., 1999; Zeeberg *et al*., 2003; Bussey *et al*., 2003). Other sources (http://www.cmb.lu.se/devbiol/bioinfo/download/intro2003/databases_part2.pdf) are in the form of flowcharts that give instructions for manual analysis.

MedMiner, GoMiner and MatchMiner can perform a synonym-based PubMed search, an ontological classification with statistical analysis of under/over-representation of differentially expressed genes in different ontological categories and a comparison of lists set in different gene identifier-formats, respectively (Tanabe *et al*., 1999; Zeeberg *et al*., 2003; Bussey *et al*., 2003). However, a user-friendly literature search based on a list of accession numbers is not available through these programs.

*To whom correspondence should be addressed.

[1]Here and further on we use accession numbers to refer to both GenBank accession numbers and UniGene cluster ids.

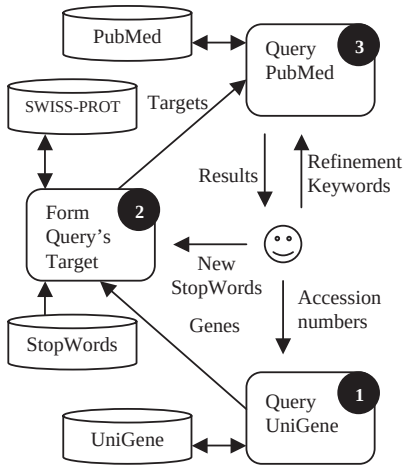

Fig. 1. MicroGENIE's data flow and architecture.

microGENIE: architecture and workflow

We set out to develop a tool specifically designed to perform an automated search of the literature based on a batch of accession numbers, allowing for a broad application spectrum. To increase efficiency, user intervention on the level of 'refinement keywords' is allowed.

Data flow and architecture of microGENIE are as depicted in Figure 1.

Information resources employed in the prototype include the UniGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene) SWISS-PROT (http://www.expasy.org/sprot/) and PubMed databases, and a list of StopWords.

From UniGene we obtained data about human genes limited to the gene name, the gene's short description and the accession numbers derived from the gene. From SWISS-PROT, we obtained only the 'GeneNames' field.

Generally, several different names are used for one gene. This ambiguity of gene names is recognized as a major problem in the field of bioinformatics (Yu and Agichtein, 2003) and most of our efforts were devoted to overcoming it by combining the information from UniGene's 'ShortDescription' and SWISS-PROTs 'GeneNames' fields.

Other tools such as MatchMiner can be used for translation from other probe identifiers to UniGene cluster ids.

Quite often, a gene's short description will include words that might aid the literature search. We obtain these words by breaking down the description into a set of terms using customary delimiters (space, comma, semicolon, etc.).

However, a gene's short description will also include terms that cannot be used in a search. We use a list of StopWords to store all irrelevant words. User intervention is required to separate the StopWords from useful keywords. Thus, a set of 'Targets' (T) is obtained, including the gene name from UniGene and synonyms from the 'GeneNames' field of SWISS-PROT.

We have considered a design with a local synonmym database to avoid repeated computation of synonyms. However, UniGene and SWISS-PROT are updated frequently and it would be difficult to synchronize these changes with such a local database.

The PubMed database contains biomedical publications and the NCBI provides a search system called Entrez (http://www.ncbi.nih.gov/Entrez/) to simplify querying. To query PubMed, Entrez-supported syntax is required, which is similar to the syntax employed by most major search engines. We use T and a set of refinement keywords (R) to form the following Boolean queries:

t_0 or t_1 or $\cdots$ or t_n;
(t_0 or t_1 or $\cdots$ or t_n) and r_0;
(t_0 or t_1 or $\cdots$ or t_n) and r_0 and r_1;
$\cdots$;
(t_0 or t_1 or . . . or t_n) and r_0 and r_1 and $\cdots$ and r_m

where $t_0 \cdots t_n$ are the components of T, and $r_0 \cdots r_m$ are the components of R (currently the user is limited to two refinement keywords).

These queries search for at least one target keyword, ranked by the presence of the number of refinement keywords.

The queries are then expressed in Entrez-supported syntax and submitted to the system. The results are summarized for the user, grouped by UniGene cluster ids (see Supplementary material for an example).

A useful feature is the listing of the number of hits both with and without the refinement keywords. This allows the user to choose the optimal balance between specificity and number of hits, without having to re-run the query. For 12 accession numbers (e.g. see Supplementary material) the refinement keywords treatment, antibody and detection reduced the average number of hits for cancer from 717 to 164, 71 and 27, respectively.

IMPLEMENTATION

MicroGENIE is implemented as a web application with a Java Servlet core. JSP technology provides a HTML interface. Xercess API handles the Entrez XML-output.

Firebird is a DBMS hosting UniGene and SWISS-PROT and is bridged to Java by JDBC technology.

Entrez imposes limitations on its usage, including a maximum of one query every 3 s. Thus, the processing of N accession numbers and M refinement keywords takes at least $3 * N * (M + 1)$ s. This is acceptable up to several hundreds accession numbers (i.e. 15 min–1 h). This limitation of Entrez is currently our main performance bottleneck, but is not inherent to our design. Even with this limitation, 15 min is acceptable for a task that is nearly impossible to do by hand.

Our software is available upon request.

DISCUSSION AND CONCLUSION

MicroGENIE proved to be very useful for performing a rapid literature search based on a list of accession numbers, as obtained from micro-array experiments (e.g. see Supplementary Material). In a typical run, we selected 200 accession numbers from micro-array data. MicroGENIE returned results within 30 min and clearly indicated frequently cited genes. From a list of accession numbers, e.g. upregulated genes, the algorithm recognized that different accession numbers represented one gene, which increased the confidence in the data for this gene.

A remaining problem is the low recall due to multiplicity in gene naming. In future work, we aim to solve this problem by including additional data sources to improve detection of synonymous gene names. Also, further studies are warranted to provide means of analyzing patterns in the text of the obtained abstracts.

REFERENCES

Bussey, K. J., Kane,D., Sunshine,M., Narasimhan,S., Nishizuka,S., Reinhold,W.C., Zeeberg,B., Ajay,W. and Weinstein,J.N. (2003) MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol.*, **4**, R27.

Tanabe,L., Scherf,U., Smith,L.H., Lee,J.K., Hunter,L. and Weinstein,J.N. (1999) MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, **27**, 1210–1214, 1216–1217.

Yu, H. and Agichtein,E. (2003) Extracting synonymous gene and protein terms from biological literature. *Bioinformatics* **19** (Suppl. 1), I340–I349.

Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T. Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C., Lababidi,S. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.